Manski, C. F., & Tetenov, A. (2016). Sufficient Trial Size to Inform Clinical Practice. Proceedings of the National Academy of Sciences of the United States of America, 113(38), 10518–10523. DOI: 10.1073/pnas.1612174113

SUFFICIENT TRIAL SIZE TO INFORM CLINICAL PRACTICE; NOT TO ACHIEVE STATISTICAL POWER

Charles F. Manski[a] and Aleksey Tetenov[b,c]

a.Department of Economics and Institute for Policy Research, Northwestern University, Evanston, IL 60208

b.Department of Economics, University of Bristol, Bristol, BS8 1TU, United Kingdom

c. Collegio Carlo Alberto, Moncalieri (TO), 10024, Italy

Abstract: Medical research has evolved conventions for choosing sample size in randomized clinical trials that rest on the theory of hypothesis testing. Bayesians have argued that trials should be designed to maximize subjective expected utility in settings of clinical interest. This perspective is compelling given a credible prior distribution on treatment response, but there is rarely consensus on what the subjective prior beliefs should be. We use Wald's frequentist statistical decision theory to study design of trials under ambiguity. We show that ε-*optimal* rules exist when trials have large enough sample size. An ε-optimal rule has expected welfare within ε of the welfare of the best treatment in every state of nature. Equivalently, it has maximum regret no larger than ε. We consider trials that draw predetermined numbers of subjects at random within groups stratified by covariates and treatments. We report exact results for the special case of two treatments and binary outcomes. We give simple sufficient conditions on sample sizes that ensure existence of ε-optimal treatment rules when there are multiple treatments and outcomes are bounded. These conditions are obtained by application of Hoeffding large deviations inequalities to evaluate the performance of empirical success rules.

Significance Statement: A core objective of trials comparing alternative medical treatments is to inform treatment choice in clinical practice. Yet conventional practice in designing trials has been to choose a sample size that yields specified statistical power. Power, a concept in the theory of hypothesis testing, is at most loosely connected to effective treatment choice. This paper develops an alternative principle for trial design that aims to directly benefit medical decision making. We propose choosing a sample size that enables implementation of near-optimal treatment rules. Near optimality means that treatment choices are suitably close to the best that could be achieved if clinicians were to know with certainty mean treatment response in their patient populations.

1. Introduction

A core objective of randomized clinical trials (RCTs) comparing alternative medical treatments is to inform treatment choice in clinical practice. Yet the conventional practice in designing trials has been to choose a sample size that yields specified statistical power. Power, a concept in the statistical theory of hypothesis testing, is at most loosely connected to effective treatment choice.

This paper develops an alternative principle for trial design that aims to directly benefit medical decision making. We propose choosing a sample size that enables implementation of near-optimal treatment rules. Near optimality means that treatment choices are suitably close to the best that could be achieved if clinicians were to know with certainty mean treatment response in their patient populations. We report exact results for the case of two treatments and binary outcomes. We derive simple formulae to compute sufficient sample sizes in clinical trials with multiple treatments.

While our immediate concern is to improve the design of RCTs, our work contributes more broadly by adding to the reasons why scientists and the general public should question the hegemony of hypothesis testing as a methodology used to collect and analyze sample data. It has become common for scientists to express concern that evaluation of empirical research by the outcome of statistical hypothesis tests generates publication bias and diminishes the reproducibility of findings. See, for example, (1) and the recent statement by the American Statistical Association (2). We call attention to a further deficiency of testing. In addition to providing an unsatisfactory basis for evaluation of research that uses sample data, testing also is deficient as a basis for the design of data collection.

2. Background

2.1. The Conventional Practice

The conventional use of statistical power calculations to set sample size in RCTs derives from the presumption that data on outcomes in a classical trial with perfect validity will be used to test a specified null hypothesis against an alternative. A common practice is to use the outcome of a hypothesis test to recommend whether a patient population should receive a status quo treatment or an innovation. The usual null hypothesis is that the innovation is no better than the status quo and the alternative is that the innovation is better. If the null hypothesis is not rejected, it is recommended that the status quo treatment continue to be used. If the null is rejected, it is recommended that the innovation replace the status quo as the treatment of choice.

The standard practice has been to perform a test that fixes the probability of rejecting the null hypothesis when it is correct, called the probability of a Type I error. Then sample size determines the probability of rejecting the alternative hypothesis when it is correct, called the probability of a Type II error. The power of a test is defined as one minus the probability of a type II error. The convention has been to choose a sample size that yields specified power at some value of the effect size deemed clinically important.

The U. S. Food and Drug Administration (FDA) uses such a test to approve new treatments. A pharmaceutical firm wanting approval of a new drug (the innovation) performs RCTs that compare the new drug with an approved drug or placebo (the status quo). An FDA document providing guidance for the design of RCTs evaluating new medical devices states that the probability of a Type I error is conventionally set to 0.05 and that the probability of a Type II error depends on the claim for the device but should not exceed 0.20 (3). The International Conference on Harmonisation (4) has provided similar guidance for the design of RCTs evaluating pharmaceuticals, stating (p. 1923): "Conventionally the probability of type I error is set at 5% or less or as dictated by any adjustments made necessary for multiplicity considerations; the precise choice may be influenced by the prior plausibility of the hypothesis under test and the desired impact of the results. The probability of type II error is conventionally set at 10% to 20%."

Trials with samples too small to achieve conventional error probabilities are called "underpowered" and are regularly criticized as scientifically useless and medically unethical. For example, Halpern, Karlawish, and Berlin (5) write (p. 358): "Because such studies may not adequately test the underlying hypotheses, they have been considered 'scientifically useless' and therefore unethical in their exposure of participants to the risks and burdens of human research." Ones with samples larger than needed to achieve conventional error probabilities are called "overpowered" and are sometimes criticized as unethical. For example, Altman (6) writes (p. 1336): "A study with an overlarge sample may be deemed unethical through the unnecessary involvement of extra subjects and the correspondingly increased costs."

2.2. Deficiencies of Using Statistical Power to Choose Sample Size

There are multiple reasons why choosing sample size to achieve specified statistical power may yield unsatisfactory results for medical decisions. These include

(1) **Use of conventional asymmetric error probabilities:** As discussed above, it has been standard to fix the probability of Type I error at 5% and the probability of Type II error for a clinically important alternative at 10-20%, which implies that the probability of Type II error reaches 95% for alternatives close to the null. The theory of hypothesis testing gives no rationale for selection of these conventional error probabilities. In particular, it gives no reason why a clinician concerned with patient welfare should find it reasonable to make treatment choices that have a substantially greater probability of Type II than Type I error.

(2) **Inattention to magnitudes of losses when errors occur:** A clinician should care about more than the probabilities of Type I and II error. He should care as well about the magnitudes of the losses to patient welfare that arise when errors occur. A given error probability should be less acceptable when the welfare difference between treatments is larger, but the theory of hypothesis testing does not take this into account.

(3) **Limitation to settings with two treatments:** A clinician often chooses among several treatments and many clinical trials compare more than two treatments. Yet the standard theory of hypothesis testing only contemplates choice between two treatments. Statisticians have struggled to extend it to deal sensibly with comparisons of multiple treatments (7, 8).

2.3. Bayesian Trial Design and Treatment Choice

With these deficiencies in mind, Bayesian statisticians have long criticized the use of hypothesis testing to design trials and make treatment decisions. The literature on Bayesian statistical inference rejects the frequentist foundations of hypothesis testing, arguing for superiority of the Bayesian practice of using sample data to transform a subjective prior distribution on treatment response into a subjective posterior distribution. See, for example, (9, 10).

The literature on Bayesian statistical decision theory additionally argues that the purpose of trials is to improve medical decision making and concludes that trials should be designed to maximize subjective expected utility in decision problems of clinical interest. The usefulness of performing a trial is expressed by the expected value of information (11), defined succinctly in Meltzer (12, p. 119) as "the change in expected utility with the collection of information." The Bayesian value of information provided by a trial crucially depends on the subjective prior distribution. The sample sizes selected in Bayesian trials may differ from those motivated by testing theory. See, for example, (13, 14).

The Bayesian perspective is compelling when a decision maker feels able to place a credible subjective prior distribution on treatment response. However, Bayesian statisticians have long struggled to provide guidance on specification of priors and the matter continues to be controversial. See, for example, the spectrum of views expressed by the authors and discussants of (9). The controversy suggests that inability to express a credible prior is common in actual decision settings.

2.4. Uniformly Satisfactory Trial Design and Treatment Choice with the Minimax-Regret Criterion

When it is difficult to place a credible subjective distribution on treatment response, a reasonable way to make treatment choices is to use a decision rule that achieves uniformly satisfactory results, whatever the true distribution of treatment response may be. There are multiple ways to formalize the idea of uniformly satisfactory results. One prominent idea motivates the minimax-regret (MMR) criterion.

Minimax regret was first suggested as a general principle for decision making under uncertainty by Savage (15) within an essay commenting on the seminal Wald (16) development of statistical decision theory. Wald considered the broad problem of using sample data to make decisions when one has incomplete knowledge of the choice environment, called the state of nature. He recommended evaluation of decision rules as procedures, specifying how a decision maker would use whatever data may be realized. In particular, he proposed measurement of the mean performance of decision rules across repetitions of the sampling process. This grounds the Wald theory in frequentist rather than Bayesian statistical thinking. See (17, 18) for comprehensive expositions.

Considering the Wald framework, Savage defined the regret associated with choice of a decision rule in a particular state of nature to be the mean loss in welfare that would occur across repeated samples if one were to choose this rule rather than the one that is best in this state of nature. The actual decision problem requires choice of a decision rule without knowing the true state of nature. The decision maker can evaluate a rule by the maximum regret that it may yield across all possible states of nature. He can then choose a rule that minimizes the value of maximum regret. Doing so yields a rule that is uniformly satisfactory in the sense of yielding the best possible upper bound on regret, whatever the true state of nature may be.

It is important to understand that maximum regret as defined by Savage is computed ex ante, before one chooses an action. It should not be confused with the familiar psychological notion of regret, which a person may perceive ex post after choosing an action and observing the true state of nature.

A decision made by the MMR criterion is invariant with respect to increasing affine transformations of welfare, but it may vary when welfare is transformed nonlinearly. The MMR criterion shares this property with expected utility maximization.

The MMR criterion is sometimes confused with the maximin criterion. A decision maker using the maximin criterion chooses an action that maximizes the minimum welfare that might possibly occur. Someone using the MMR criterion chooses an action that minimizes the maximum loss to welfare that can possibly result from not knowing the welfare function. Whereas the maximin criterion considers only the worst outcome that an action may yield, MMR considers the worst outcome relative to what is achievable in a given state of nature. Savage (15), while introducing the MMR criterion, distinguished it sharply from maximin, writing that the latter criterion is "ultrapessimistic" while the former is not.

Since the early 2000s, various authors have used the MMR criterion to study how a decision maker might use RCT data to subsequently choose treatments for the members of a population (19-27). In these studies, the decision maker's objective has been expressed as maximization of a welfare function that sums treatment outcomes across the population. For example, the objective may be to maximize the five-year survival rate of a population of cancer patients or the average number of quality-adjusted life years of a population with a chronic disease.

The MMR criterion is applicable in general settings with multiple treatments. Regret is easiest to explain when there are two treatments, say A and B. If treatment A is better, regret is the probability of a Type I error (choosing B) times the magnitude of the resulting loss in population welfare due to assigning the inferior treatment. Symmetrically, if treatment B is better, regret is the probability of a Type II error (choosing A) times the magnitude of the resulting loss in population welfare due to foregoing the superior treatment. In contrast to the use of hypothesis testing to choose a treatment, the MMR criterion views Type I and II error probabilities symmetrically and it assesses the magnitudes of the losses that errors produce.

Whereas the work cited above has used the MMR criterion to guide treatment choice after a trial has been performed, the present paper uses it to guide the design of RCTs. We focus on classical trials possessing perfect validity that compare alternative treatments relevant to clinical practice. Treatments may include placebo if it is a relevant clinical option or if it is considered equivalent to prescribing no treatment (28, 29). In particular, we study trials that draw subjects at random within groups of predetermined size stratified by covariates and treatments. Section 3 summarizes the major findings. The online

Supporting Information section provides underlying technical analysis.

3. Trials Enabling Near-Optimal Treatment Rules

3.1. General Ideas

An ideal objective for trial design would be to collect data that enable subsequent implementation of an optimal treatment rule in a population of interest−−one that always selects the best treatment, with no chance of error. Optimality is too strong a property to be achievable with trials having finite sample size, but *near-optimal* rules exist when classical trials with perfect validity have large enough size.

Given a specified $\varepsilon > 0$, an ε-optimal rule is one whose mean performance across samples is within ε of the welfare of the best treatment, whatever the true state of nature may be. Equivalently, an ε-optimal rule has maximum regret no larger than ε. Thus, an ε-optimal rule exists if and only if the MMR rule has maximum regret no larger than ε.

Choosing sample size to enable existence of ε-optimal treatment rules provides an appealing criterion for design of trials that aim to inform treatment choice. Implementation of the idea requires specification of a value for ε. The need to choose an effect size of interest when designing trials already arises in conventional practice, where the trial planner must specify the alternative hypothesis to be compared with the null. A possible way to specify ε is to make it equal the *minimum clinically important difference* (MCID) in the average treatment effect comparing alternative treatments.

Medical research has long distinguished between the statistical and clinical significance of treatment effects (30). While the idea of clinical significance has been interpreted in various ways, many writers call an average treatment effect clinically significant if its magnitude is greater than a specified value deemed minimally consequential in clinical practice. The ICH (4) put it this way (p. 1923): "The treatment difference to be detected may be based on a judgment concerning the minimal effect which has clinical relevance in the management of patients."

Research articles reporting trial findings sometimes pose particular values of MCIDs when comparing alternative treatments for specific diseases. For example, in a study comparing drug treatments for hypertension, Materson *et al.* (31) defined the outcome of interest to be the fraction of subjects who achieve a specified threshold for blood pressure. They took the MCID to be the fraction 0.15, stating that this is (p. 916): "the difference specified in the study design to be clinically important" and reported groups of drugs "whose effects do not differ from each other by more than 15 percent."

3.2. Findings with Binary Outcomes, Two Treatments, and Balanced Designs

Determination of sample sizes that enable near-optimal treatment is simple in settings with binary outcomes (coded 0 and 1 for simplicity), two treatments, and a balanced design which assigns the same number of subjects to each treatment group. Table 1 provides exact computations of the minimum sample size that enables ε-optimality when a clinician uses one of three different treatment rules, for various values of ε.

The first column shows the minimum sample size (per treatment arm) that yields ε-optimality when a clinician uses the empirical success (ES) rule to make a treatment decision. The ES rule chooses the treatment with the better average outcome in the trial. The rule assigns half the population to each treatment if there is a tie. It is known that the ES rule minimizes maximum regret rule in settings with binary outcomes, two treatments, and balanced designs (25).

The second and third columns display the minimum sample sizes that yield ε-optimality of rules based on one-sided 5% and 1% hypothesis tests. There is no consensus on what hypothesis test should be used to compare two proportions. We report results based on the widely used one-sided two-sample z-test, which is based on an asymptotic normal approximation (32).

The findings are remarkable. A sample as small as 2 observations per treatment arm makes the ES rule ε-optimal when $\varepsilon = 0.1$ and a sample of size 145 suffices when $\varepsilon = 0.01$. The minimum sample sizes required for ε-optimality of the test rules are orders of magnitude larger. If the z-test of size 0.05 is used, a sample of size 33 is required when $\varepsilon = 0.1$ and 3488 when $\varepsilon = 0.01$. The sample sizes have to be more than double these values if the z-test of size 0.01 is used.

Figure 1 illustrates the difference between error probabilities and regret incurred by the ES rule and the 5% z-test rule for a sample size of 145 per arm, the minimum sample size yielding ε-optimality when $\varepsilon = 0.01$. The top panels show how the probability of error varies with the effect size for all possible distributions of treatment response with effect sizes in the range [-0.5, 0.5]. The bottom panels display the regret (probability of error times the effect size) of the same treatment rules. Maximum regret occurs at intermediate effect sizes. For small effect sizes, regret is small because choosing the wrong treatment is not clinically significant. Regret is also small for large effect sizes, because the probability of error eventually starts declining rapidly with the effect size. Traditional power calculations are not informative about the maximum regret of a test-based rule. Two red vertical lines mark effect sizes at which the z-test has at least 80% and 90% power. Neither corresponds to the effect size where regret is maximal.

3.3. Findings with Bounded Outcomes and Multiple Treatments

In principle, the existence of ε-optimal treatment rules under any design can be determined by computing the maximum regret of the minimax-regret rule. In practice, determination of the minimax-regret rule and its maximum regret may be burdensome. To date, exact minimax-regret decision rules have been derived only for the case of two treatments with equal or nearly-equal sample sizes (24-26). Hence, it is useful to have simple sufficient conditions that ensure existence of ε-optimal rules more generally. The conditions we derive below hold in all settings where outcomes are bounded. Our findings apply to situations in which there are multiple treatments, not just two. They also apply when trials stratify patients into groups with different observable covariates, such as demographic attributes and risk factors.

To show that a specified trial design enables ε-optimal treatment rules, it suffices to consider a particular rule and to show that this rule is ε-optimal when used with this design. We focus on empirical success rules for both practical and analytical reasons. Choosing a treatment with the highest reported mean outcome is a simple and plausible way in which a clinician may use the results of an RCT. Two analytical reasons further motivate interest in ES rules when outcomes are bounded. First, these rules either exactly or approximately minimize maximum regret in various settings with two treatments when sample size is moderate (25, 26) and asymptotically (23). Second, large deviations inequalities derived in (33) allow us to obtain informative and easily computable upper bounds on the maximum regret of ES rules applied with any number of treatments. These upper bounds on maximum regret immediately yield sample sizes that ensure an ES rule is ε-optimal.

Propositions 1 and 2 (see the Supporting Information) present two alternative upper bounds on the maximum regret of an ES rule. Proposition 1 extends finding of Manski (19) from two to multiple treatments while Proposition 2 derives a new large-deviations bound for multiple treatments. When the design is balanced, these bounds are

(1) $(2e)^{-½}M(K-1)n^{-½}$,

(2) $M(\ln K)^{½}n^{-½}$,

where n is the sample size per arm, K is the number of treatment arms, and M is the width of the range of possible outcomes. Proposition 3 (see the Supporting Information) shows that the bounds on maximum regret derived in Propositions 1 and 2 are minimized by balanced

designs. Section 3.4 extends these findings to settings where patients have observable covariates.

Propositions 1 and 2 imply sufficient conditions on sample sizes for ε-optimality of ES rules. Proposition 1 implies that an ES rule is ε-optimal if the sample size per treatment arm is at least

(3) $n \geq (2e)^{-1}(K-1)^2(M/\varepsilon)^2$.

Proposition 2 implies that an ES rule is ε-optimal if the sample size per treatment arm is at least

(4) $n \geq \ln K\,(M/\varepsilon)^2$.

We find that when the design is balanced, Proposition 1 provides a tighter bound than Proposition 2 for two or three treatments. Proposition 2 gives a tighter bound for four or more treatments.

To illustrate the findings, consider the Materson *et al.* (31) study of treatment for hypertension described in the Introduction. The outcome is binary with the range of possible outcomes $M = 1$. The study compared seven drug treatments and specified 0.15 as the MCID. We cannot know how the authors of the study, who reported results of traditional hypothesis tests, would have specified ε had they sought to achieve ε-optimality. If they were to set $\varepsilon = 0.15$, application of bound (4) shows that an ES rule is ε-optimal if the number of subjects per treatment arm is at least $(\ln 7)\cdot(0.15)^{-2} = 86.5$. The actual study has an approximately balanced design, with between 178 and 188 subjects in each treatment arm. Application of bound (2) shows that a study with at least 178 subjects per arm is ε-optimal for $\varepsilon = (\ln 7)^{½}(178)^{-½} = 0.105$.

It is important to bear in mind that Propositions 1 and 2 only imply simple sufficient conditions on sample sizes for ε-optimality of ES rules, not necessary ones. These sufficient conditions use only the weak assumption that outcomes are bounded and they rely on Hoeffding large-deviations inequalities for bounded outcomes. In the special case of Section 3.1---with binary outcomes, two treatments, and a balanced design--the sufficient sample sizes provided by Proposition 1 are roughly ten times the size of the exact minimum sample sizes, depending on the value of ε. This strongly suggests that it is worthwhile to compute exact minimum sample sizes whenever it is tractable to do so.

3.4. Trials Stratified by Observed Covariates

Clinical trials often stratify participants by observable covariates, such as demographic attributes and risk factors, and report trial results separately for each group. We consider ε-optimality of the ES rule which assigns individuals with covariates ξ to the treatment which yielded the highest average outcome among trial participants with covariates ξ.

There are at least two reasonable ways that a planner may wish to evaluate ε-optimality in this setting. First, he may want to achieve ε-optimality within each covariate group. This interpretation requires no new analysis. The planner should simply define each covariate group to be a separate population of interest and then apply the analysis of Sections 3.2–3.3 to each group. The design that achieves group-specific ε-optimality with minimum total sample size equalizes sample sizes across groups.

Alternatively, the planner may want to achieve ε-optimality within the overall population, without requiring that it be achieved within each covariate group. Bounds (1) and (2) extend to the setting with covariates. With a balanced design assigning n_ξ individuals from covariate group ξ to each treatment, the maximum regret of an ES rule is bounded above by

(5) $(2e)^{-½}\,M(K-1)\sum_{\xi \in X} P(x=\xi)(n_\xi)^{-½}$,

(6) $M\,(\ln K)^{½}\sum_{\xi \in X} P(x=\xi)(n_\xi)^{-½}$.

The design that minimizes bound (5) or (6) for a given total sample size generally neither equalizes sample sizes across groups, nor makes them proportional to the covariate distribution $P(x = \xi)$. Instead, the relative sample sizes for any pair (ξ, ξ') of covariate values have the approximate ratio

(7) $n_\xi/n_{\xi'} = [P(x=\xi)/P(x=\xi')]^{⅔}$.

Such trial designs make the covariate-specific sample size increase with the prevalence of the covariate group in the population, but less than proportionately. Covariate-specific maximum regret commensurately decreases with the prevalence of the covariate group.

4. Conclusion

Choosing sample sizes in clinical trials to enable near-optimal treatment rules would align trial design directly with the objective of informing treatment choice. In contrast, the conventional practice of choosing sample size to achieve specified statistical power in hypothesis testing is only loosely related to treatment choice. Our work adds to the growing concern of scientists that hypothesis testing provides an unsuitable methodology for collection and analysis of sample data.

We share with Bayesian statisticians who have written on trial design the objective of informing treatment choice. We differ in our application of the frequentist statistical decision theory developed by Wald, which does not require that one place a subjective probability distribution on treatment response. We use the concept of ε-optimality, which is equivalent to having maximum regret no larger than ε.

There are numerous potentially fruitful directions for further research of the type initiated here. One is analysis of other types of trials. We have focused on trials that draw subjects at random within groups of predetermined size stratified by covariates and treatments. With further work, the ideas developed here should be applicable to trials where the numbers of subjects who have particular covariates and receive specific treatment are ex ante random rather than predetermined.

Our analysis assumed no prior knowledge restricting the variation of response across treatments and covariates. This assumption, which has been traditional in frequentist study of clinical trials, is advantageous in the sense that it yields generally applicable findings. Nevertheless, it is unduly conservative in circumstances where some credible knowledge of treatment response is available. One may, for example, think it credible to maintain some assumptions on the degree to which treatment response may vary across treatments or covariate groups. When such assumptions are warranted, it may be valuable to impose them.

We mentioned at the outset that medical conventions for choosing sample size pertain to classical trials possessing perfect validity. However, practical trials usually have only partial validity. For example, the experimental sample may be representative only of a part of the target treatment population, because experimental subjects typically are persons who meet specified criteria and who consent to participate in the trial. Due to this and other reasons, experimental data may only partially identify treatment response in the target treatment population. The concept of ε-optimality extends to such situations.

Finally, we remark that our analysis followed the longstanding practice in medical research of evaluating trial designs by their informativeness about treatment response, without consideration of the cost of conducting trials. The concept of ε-optimality can be extended to recognize trial cost as a determinant of welfare.

Acknowledgements: We have benefitted from the comments of Joerg Stoye and from the opportunity to present this work in a seminar at Cornell University.

References

1. Ioannidis J (2005) Why most published research findings are false. *PLoS Med.* 2, e124.
2. Wasserstein R, Lazar N (2016) The ASA's Statement on p-Values: Context, Process, and Purpose. *Am Stat* 70(2):129-133.
3. U.S. Food and Drug Administration (1996) *Statistical Guidance for Clinical Trials of Nondiagnostic Medical Devices.* http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm106757.htm.
4. International Conference on Harmonisation (1999) ICH E9 Expert Working Group. Statistical principles for clinical trials: ICH harmonized tripartite guideline. *Stat. Med.* 18:1905-1942.
5. Halpern S, Karlawish J, Berlin J (2002) The continued unethical conduct of underpowered clinical trials. *JAMA* 288:358-362.
6. Altman D (1980) Statistics and ethics in medical research: III How large a sample? *BMJ* 281: 1336-1338.
7. Dunnett C (1955) A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *JASA* 50: 1096-1121.
8. Cook R, Farewell V (1996) Multiplicity Considerations in the Design and Analysis of Clinical Trials. *J R Stat Soc Series A* 159:93-110.
9. Spiegelhalter D, Freedman L, Parmar M (1994) Bayesian approaches to randomized trials (with discussion). *J R Stat Soc Series A* 157:357-416.
10. Spiegelhalter D (2004) Incorporating Bayesian ideas into health-care evaluation. *Stat Sci* 19: 156-174.
11. Claxton K, Posnett J (1996) An economic approach to clinical trial design and research priority-setting. *J Health Econ* 5:513-524.
12. Meltzer D (2001) Addressing uncertainty in medical cost-effectiveness: analysis implications of expected utility maximization for methods to perform sensitivity analysis and the use of cost-effectiveness analysis to set priorities for medical research. *J Health Econ* 20(1):109–129.
13. Cheng Y, Su F, Berry D (2003) Choosing sample size for a clinical trial using decision analysis. *Biometrika* 90:923-936.
14. Berry D (2004) Bayesian statistics and the efficiency and ethics of clinical trials. *Stat Sci* 19:175-187.
15. Savage L (1951) The theory of statistical decision. *JASA* 46:55-67.
16. Wald A (1950) *Statistical Decision Functions* (Wiley, New York).
17. Ferguson T (1967) *Mathematical Statistics: A Decision Theoretic Approach* (Academic Press, San Diego).
18. Berger J (1985) *Statistical Decision Theory and Bayesian Analysis*, Second Edition (Springer, New York).
19. Manski C (2004) Statistical treatment rules for heterogeneous populations. *Econometrica* 72:221-246.
20. Manski C (2005) *Social Choice with Partial Knowledge of Treatment Response* (Princeton Univ. Press, Princeton).
21. Manski C (2007) Minimax-regret treatment choice with missing outcome data. *J Econometrics* 139:105-115.
22. Manski C, Tetenov A (2007) Admissible treatment rules for a risk-averse planner with experimental data on an innovation. *J Stat Plan* 137:1998-2010.
23. Hirano K, Porter J (2009) Asymptotics for statistical treatment rules. *Econometrica* 77:1683-1701.
24. Schlag K (2006) "Eleven – tests needed for a recommendation" (EUI Working Paper ECO No. 2006/2; http://hdl.handle.net/1814/3937).
25. Stoye J (2009) Minimax regret treatment choice with finite samples. *J Econometrics* 151:70-81.
26. Stoye J (2012) Minimax regret treatment choice with covariates or with limited validity of experiments. *J Econometrics* 166:138-156.
27. Tetenov A (2012) Statistical treatment choice based on asymmetric minimax regret criteria. *J Econometrics* 166:157-165.
28. Hróbjartsson A, Gøtzsche P (2001) Is the Placebo Powerless? — An Analysis of Clinical Trials Comparing Placebo with No Treatment. *N Engl J Med* 344:1594-1602.
29. Lichtenberg P, Heresco-Levy U, Nitzan U (2004) The ethics of the placebo in clinical practice. *J Med Ethics* 30:551-554.
30. Sedgwick P (2014) Clinical significance versus statistical significance. *BMJ*, 348:g2130, doi: 10.1136/bmj.g2130.
31. Materson B et al. (1993) Single-drug therapy for hypertension in men: A comparison of six antihypertensive agents with placebo. *N Engl J Med* 328:914-921.
32. Fleiss J (1973) *Statistical Methods for Rates and Proportions* (Wiley, New York).
33. Hoeffding W (1963) Probability inequalities for sums of bounded random variables. *JASA* 58:13-30.
34. Lugosi G (2002) Pattern Classification and Learning Theory. *Principles of Nonparametric Learning*, ed. Gyorfi L (Springer, Vienna), pp. 1-56.
35. Bentkus V (2004) On Hoeffding's inequalities. *Ann Prob* 32: 1650-1673.

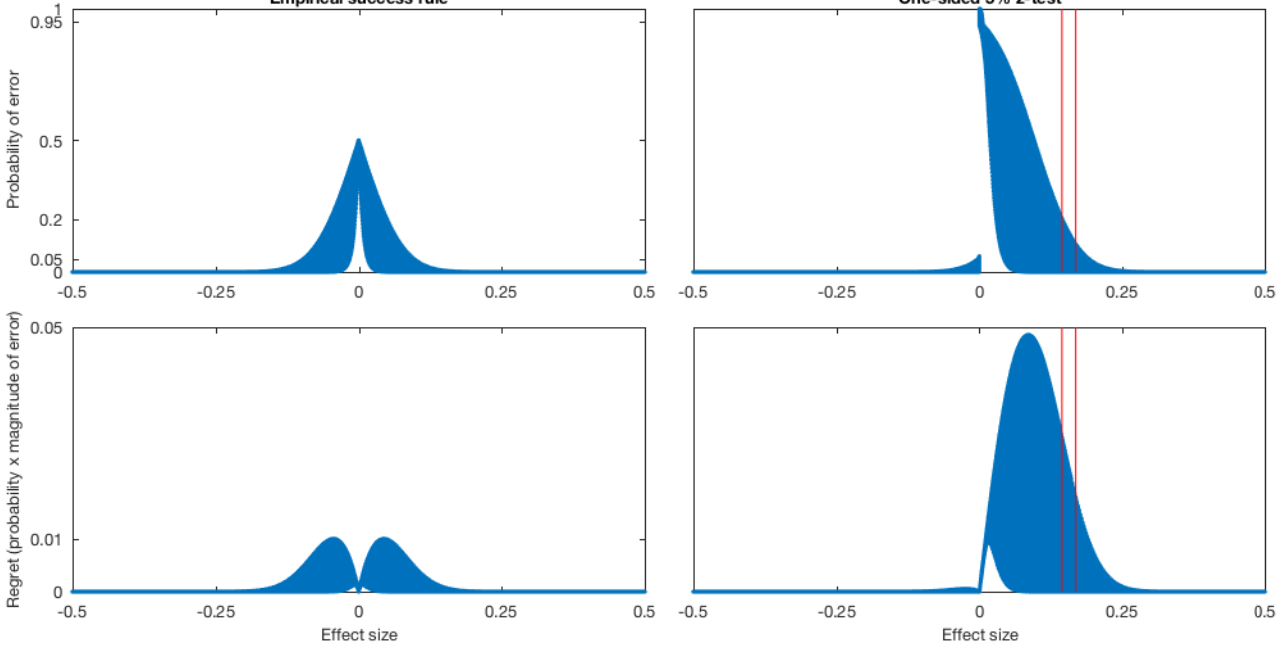
Empirical success rule
One-sided 5% z-test
Probability of error
1
0.95
0.5
0.2
0.05
0
-0.5
-0.25
0
0.25
0.5
Regret (probability x magnitude of error)
0.05
0.01
0
Effect size

Table 1: Minimum Sample Sizes per Treatment Enabling ε-Optimal Treatment Choice: Binary Outcomes, Two Treatments, Balanced Designs

ε	ES Rule	One-Sided 5% z-Test	One-Sided 1% z-Test
0.01	145	3488	7963
0.03	17	382	879
0.05	6	138	310
0.10	2	33	79
0.15	1	16	35

SUFFICIENT TRIAL SIZE TO INFORM CLINICAL PRACTICE; NOT TO ACHIEVE STATISTICAL POWER

SUPPORTING INFORMATION

S1. Principles for Evaluation of Trial Designs and Treatment Rules

S1.1. The Decision Problem

The setup is as in (19, 22). A planner must assign one of K treatments to each member of a treatment population, denoted J. Let T denote a finite set of feasible treatments. Each $j \in J$ has a response function $u_j(\cdot)$: $T \to U$ mapping treatments $t \in T$ into individual welfare outcomes $u_j(t) \in \mathbf{R}$. Treatment is individualistic; that is, a person's outcome may depend on the treatment he is assigned but not on the treatments assigned to others. The population is a probability space (J, Ω, P), and the probability distribution $P[u(\cdot)]$ of the random function $u(\cdot)$: $T \to \mathbf{R}$ describes treatment response across the population. The population is "large;" formally J is uncountable and $P(j) = 0, j \in J$.

Let person j have observable covariates x_j taking a value in a covariate space X; thus, x: $J \to X$ is the random variable mapping persons into their covariates. We suppose that X is finite with $P(x = \xi) > 0, \ \forall\ \xi \in X$. We also suppose that the covariate distribution $P(x)$ is known. The planner can systematically differentiate persons with different observed covariates, but he cannot distinguish among persons with the same observed covariates.

A statistical treatment rule maps sample data into a treatment allocation. Let Q denote the sampling distribution generating the available data and let Ψ denote the sample space; that is, Ψ is the set of data samples that may be drawn under Q. A feasible treatment rule is a function that assigns all persons with the same observed covariates to one treatment or, more generally, a function that randomly allocates such persons across the different treatments. Let Δ now denote the space of functions that map $T \times X \times \Psi$ into the unit interval and that satisfy the adding-up conditions: $\delta \in \Delta \ \Rightarrow\ \sum_{t \in T} \delta(t, \xi, \psi) = 1, \ \forall\ (\xi, \psi) \ \in X \times \Psi$. Then each function $\delta \in \Delta$ defines a statistical treatment rule.

The planner wants to maximize population welfare, which adds welfare outcomes across persons. Given data ψ, the population welfare that would be realized if the planner were to choose rule δ is

$$\text{(S1)} \quad U(\delta, P, \psi) \ \equiv \ \sum_{\xi \in X} P(x = \xi) \sum_{t \in T} \delta(t, \xi, \psi) \cdot \mathrm{E}[u(t) | x = \xi] \ = \ \sum_{\xi \in X} P(x = \xi) \sum_{t \in T} \delta(t, \xi, \psi) \cdot \mu_{t\xi} \,,$$

where $\mu_{t\xi} \equiv \mathrm{E}[u(t)|x = \xi]$ is the mean outcome of treatment t among individuals with covariates ξ. Inspection of (S1) shows that, whatever value ψ may take, it is optimal to set $\delta(t, \xi, \psi) = 0$ if $\mu_{t\xi} < \max_{t' \in T} \mu_{t'\xi}$.

The problem of interest is treatment choice when one does not have enough knowledge of $P[u(\cdot)|x]$ to determine the optimal treatment that maximizes $\mu_{t\xi}$ for each $\xi \in X$.

S1.2. Evaluating Treatment Rules by their State-Dependent Welfare Distributions

The starting point for development of implementable criteria for treatment choice under uncertainty is specification of a state space, say S. Thus, let $\{(P_s, Q_s), s \in S\}$ be the set of (P, Q) pairs that the planner deems possible.

Considered as a function of ψ, $U(\delta, P_s, \psi)$ is a random variable with state-dependent sampling distribution $Q_s[U(\delta, P_s, \psi)]$. Following Wald's view of statistical decision functions as procedures, we use the vector $\{Q_s[U(\delta, P_s, \psi)], s \in S\}$ of state-dependent welfare distributions to evaluate rule δ. In principle this vector is computable, whatever the state space and sampling process may be. Hence, in principle, a planner can compare the vectors of state-dependent welfare distributions yielded by different STRs and base treatment choice on this comparison.

How might a planner compare the state-dependent welfare distributions yielded by different STRs? The planner wants to maximize welfare, so it seems self-evident that he should weakly prefer rule δ to an alternative

rule δ' if, in every $s \in S$, $Q_s[U(\delta, P_s, \psi)]$ equals or stochastically dominates $Q_s[U(\delta', P_s, \psi)]$. It is less obvious how one should compare rules whose state-dependent welfare distributions are not uniformly ordered in this manner, as is typically the case.

Wald evaluated statistical decision functions by their mean performance across realizations of the sampling process and this has become the standard practice in the subsequent literature. The expected welfare yielded by rule δ in state s, denoted $W(\delta, P_s, Q_s)$, is

$$(S2) \quad W(\delta, P_s, Q_s) \equiv \int_\Psi \sum_{\xi \in X} P(x = \xi) \sum_{t \in T} \delta(t, \xi, \psi) \cdot \mu_{st\xi} \, dQ_s(\psi) = \sum_{\xi \in X} P(x = \xi) \sum_{t \in T} \mathrm{E}_s[\delta(t, \xi, \psi)] \cdot \mu_{st\xi}.$$

Here $\mathrm{E}_s[\delta(t, \xi, \psi)] \equiv \int_\Psi \delta(t, \xi, \psi) dQ_s(\psi)$ is the expected (across potential samples) fraction of persons with covariates ξ who are assigned to treatment t. We add subscript s to $\mu_{st\xi}$ because mean treatment response varies across $s \in S$.

S1.3. Optimality and ε-Optimality of Treatment Rules

A planner must confront the fact that the true state of nature is unknown. The maximum welfare achievable in each state s is

$$(S3) \quad U^*(P_s) \equiv \sum_{\xi \in X} P(x = \xi) \cdot \max_{t \in T} \mu_{st\xi}.$$

We define rule δ to be *mean optimal* if $W(\delta, P_s, Q_s) = U^*(P_s)$ for all $s \in S$. Mean optimality is desirable, but it is too strong to be achievable in general. The concept of mean ε-optimality relaxes mean optimality, yielding a property that may be achievable in practice.

We define rule δ to be *mean ε-optimal* for a specified $\varepsilon > 0$ if $W(\delta, P_s, Q_s) \geq U^*(P_s) - \varepsilon$ for all $s \in S$. Section S2 shows that mean ε-optimal treatment rules exist when treatment outcomes are bounded and classical trials have sufficient finite size, whatever the state space may be. This finding makes ε-optimality a practical criterion for trial design.

Stating that an STR is ε-optimal is equivalent to stating that it has maximum regret no larger than ε. By definition, the regret of rule δ in state s is $U^*(P_s) - W(\delta, P_s, Q_s)$. The maximum regret of δ across all states is $\max_{s \in S} [U^*(P_s) - W(\delta, P_s, Q_s)]$. Thus, maximum regret is less than or equal to ε if and only if $U^*(P_s) - W(\delta, P_s, Q_s) \leq \varepsilon$ for all $s \in S$. It follows that mean ε-optimal STRs exist with a specified design if and only if the maximum regret of the minimax-regret rule is less than or equal to ε.

S2. Randomized Trials with Sample Sizes Enabling ε-Optimal Treatment

We now investigate the existence of ε-optimal treatment rules when the data are generated by classical randomized trials. We specifically consider trials that draw subjects at random within groups stratified by covariates and treatments. Thus, for $(t, \xi) \in T \times X$, the experimenter draws $n_{t\xi}$ subjects at random from the subpopulation with covariates ξ and assigns them to treatment t. The set $n_{TX} \equiv [n_{t\xi}, (t, \xi) \in T \times X]$ of stratum sample sizes defines the design. Let $N(t, \xi)$ be the realized sample of subjects with covariates ξ who are assigned to treatment t. The data are the sample outcomes $\psi = [u_j, j \in N(t, \xi); (t, \xi) \in T \times X]$. We suppose throughout that the state space S contains all distributions of treatment response. Thus, the planner has no prior knowledge restricting the variation of response across treatments and covariates.

In principle, the existence of ε-optimal STRs under any design can be determined by computing the maximum regret of the minimax-regret (MMR) rule. As noted earlier, ε-optimal rules exist if and only if the MMR rule has maximum regret less than or equal to ε. In practice, determination of the MMR rule and computation of its maximum regret may be burdensome. Hence, it is useful to have simple sufficient conditions that ensure existence of ε-optimal rules. This section provides such conditions in settings where outcomes are bounded.

S2.1. Sufficient Conditions for ε-Optimality of Empirical Success Rules

To show that a specified trial design enables ε-optimal STRs, it suffices to consider a particular STR and to show that this rule is ε-optimal when used with this design. We focus on *empirical success (ES)* rules, which use the empirical distribution of the sample data to estimate the population distribution of treatment response. Formally, let $m_{t\xi}(\psi)$ be the average outcome in treatment-covariate sub-sample $N(t, \xi)$; that is, $m_{t\xi}(\psi) \equiv (1/n_{t\xi})\sum_{j \in N(t, \xi)} u_j$. An ES rule δ assigns all persons with covariates ξ to treatments that maximize $m_{t\xi}(\psi)$ over T. Thus, $\delta(t, \xi, \psi) = 0$ if $m_{t\xi}(\psi) < \max_{t' \in T} m_{t'\xi}(\psi)$.

The case of two treatments has been studied previously in Manski (19), who exploited the large deviations result of Hoeffding (33) to derive an upper bound on the maximum regret of a class of ES rules that condition treatment on alternative subsets of the observable covariates of population members. The bound takes a particularly simple form when one conditions on all observable covariates and the state space includes all distributions of treatment response.

Let outcomes lie in the bounded range $[u_l, u_h]$, whose width we denote by $M \equiv u_h - u_l$. Label the two treatments $t = a$ and $t = b$, and let S index all distributions of treatment response. Manski (19) showed in eq. 23 that the maximum regret of an ES rule δ is bounded from above as follows:

$$\text{(S4)} \quad \max_{s \in S} [U^*(P_s) - W(\delta, P_s, Q_s)] \leq \tfrac{1}{2} e^{-\frac{1}{2}} M \sum_{\xi \in X} P(x = \xi)(n_{a\xi}^{-1} + n_{b\xi}^{-1})^{\frac{1}{2}}.$$

Hence, an ES rule is ε-optimal if the trial sample sizes satisfy the inequality

$$\text{(S5)} \quad \tfrac{1}{2} e^{-\frac{1}{2}} M \sum_{\xi \in X} P(x = \xi)(n_{a\xi}^{-1} + n_{b\xi}^{-1})^{\frac{1}{2}} \leq \varepsilon.$$

When the design is balanced, with $n_{t\xi} = n$ for all (t, ξ), inequality (S5) reduces to $(2e)^{-\frac{1}{2}} M n^{-\frac{1}{2}} \leq \varepsilon$. Hence, an ES rule with a balanced design is ε-optimal if $n \geq (2e)^{-1}(M/\varepsilon)^2$.

In what follows we present new findings that hold with any finite number K of treatments. Section S2.2 considers trial design when members of the population have no observable covariates. Section S2.3 extends the analysis to settings with covariates.

S2.2. Large Deviation Bounds on Maximum Regret with Multiple Treatments

Propositions 1 and 2 present two alternative upper bounds on the maximum regret of an ES rule. Proposition 1 extends inequality (S4) to multiple treatments while Proposition 2 derives a different type of bound. We find that when the design is balanced, with $n_t = n$ for all t, Proposition 1 provides a tighter bound than Proposition 2 when there are two or three treatments. Proposition 2 gives a tighter bound when there are four or more treatments. Proposition 3 shows that, for any given total sample size that is an integer multiple of K, the bounds on maximum regret derived in Propositions 1 and 2 are minimized by balanced designs.

In all propositions, a design is a vector of sample sizes $(n_t, t \in T)$ and $t^* \in \text{argmin}_{t \in T} n_t$ denotes a treatment with the smallest sample size. In the proofs we let $t(s)$ designate any one of the optimal treatments in state s; that is, $\mu_{st(s)} \geq \mu_{st}$ for all $t \in T$.

Proposition 1: The maximum regret of an empirical success rule δ is bounded above as follows:

(S6) $\max_{s \in S} [U^*(P_s) - W(\delta, P_s, Q_s)] \leq \frac{1}{2} e^{-\frac{1}{2}} M \sum_{t \neq t^*} (n_t^{-1} + n_{t^*}^{-1})^{\frac{1}{2}}$.

When the design is balanced, with $n_t = n$ for all t, the bound is $(2e)^{-\frac{1}{2}} M (K - 1) n^{-\frac{1}{2}}$. □

Proof: Given that δ is an empirical success rule, $\delta(t, \psi) \leq 1[m_t \geq m_{t'}]$ for all (t, t') in T and all $\psi \in \Psi$. Therefore, $E_s[\delta(t, \psi)] \leq P_s(m_t \geq m_{t'})$ in each state s. Hence, $E_s[\delta(t, \psi)] \leq P_s(m_t \geq m_{t(s)})$. The best achievable welfare in state s is $U^*(P_s) = \mu_{st(s)}$. Hence, the regret of δ in state s is

$$U^*(P_s) - W(\delta, P_s, Q_s) = \mu_{st(s)} - \sum_{t \in T} E_s[\delta(t, \psi)]\mu_{st} = \sum_{t \neq t(s)} E_s[\delta(t, \psi)](\mu_{st(s)} - \mu_{st}) \leq$$

$$\leq \sum_{t \neq t(s)} (\mu_{st(s)} - \mu_{st}) \cdot P_s(m_t \geq m_{t(s)}).$$

Adaptation of the argument used by Manski (19) to obtain inequality (S4) from Hoeffding's Theorem 2 (33) shows that

$$(\mu_{st(s)} - \mu_{st}) \cdot P(m_t \geq m_{t(s)}) \leq \frac{1}{2} e^{-\frac{1}{2}} M (n_t^{-1} + n_{t(s)}^{-1})^{\frac{1}{2}}.$$

It follows that

$$U^*(P_s) - W(\delta, P_s, Q_s) \leq \frac{1}{2} e^{-\frac{1}{2}} M \sum_{t \neq t(s)} (n_t^{-1} + n_{t(s)}^{-1})^{\frac{1}{2}}.$$

Hence, maximum regret is bounded above as follows:

$$\max_{s \in S} [U^*(P_s) - W(\delta, P_s, Q_s)] \leq \frac{1}{2} e^{-\frac{1}{2}} M \max_{s \in S} \sum_{t \neq t(s)} (n_t^{-1} + n_{t(s)}^{-1})^{\frac{1}{2}}.$$

Finally, the summation $\sum_{t \neq t(s)} (n_t^{-1} + n_{t(s)}^{-1})^{\frac{1}{2}}$ is maximized in a state s such that $t(s) = t^*$, where t^* is a treatment with the smallest sample size. This holds because $n_{t^{**}} \geq n_{t^*}$ for any $t^{**} \neq t^*$. Hence,

$$\sum_{t \neq t^*} (n_t^{-1} + n_{t^*}^{-1})^{\frac{1}{2}} - \sum_{t \neq t^{**}} (n_t^{-1} + n_{t^{**}}^{-1})^{\frac{1}{2}} = \sum_{t \neq t^*, t^{**}} [(n_t^{-1} + n_{t^*}^{-1})^{\frac{1}{2}} - (n_t^{-1} + n_{t^{**}}^{-1})^{\frac{1}{2}}] \geq 0.$$

Thus, (S6) holds.

Q. E. D.

Proposition 2: The maximum regret of an empirical success rule δ is bounded above by

(S7) $$\max_{s \in S} [U^*(P_s) - W(\delta, P_s, Q_s)] \leq N^{-\frac{1}{2}} M \min_{d > 0} \frac{\ln\{1 + \sum_{t \neq t^*} \exp[d^2 (p_t^{-1} + p_{t^*}^{-1})/8]\}}{d},$$

where $N \equiv \sum_{t \in T} n_t$ is total sample size and $p_t \equiv n_t / N$. When the design is balanced, with $n_t = n$ for all t, (S7) implies the bound

(S8) $\max_{s \in S} [U^*(P_s) - W(\delta, P_s, Q_s)] \leq M (\ln K)^{\frac{1}{2}} n^{-\frac{1}{2}}$. □

Proof: The proof of (S7) is in four parts. Result (S8) is then proved in Part V.

I: Fix state s and consider a treatment $t(s)$ that is optimal in this state. Fix the sample data ψ. Let

$$D_{s[t,\,t(s)]}(\psi) \equiv [m_t(\psi) - m_{t(s)}(\psi)] - (\mu_{st} - \mu_{st(s)})$$

denote the amount by which $m_t(\psi) - m_{t(s)}(\psi)$ overestimates $\mu_{st} - \mu_{st(s)}$. Note that $D_{s[t(s),t(s)]}(\psi) = 0$. We first show that the welfare loss $U^*(P_s) - U(\delta, P_s, \psi)$ is bounded above by

$$U^*(P_s) - U(\delta, P_s, \psi) \leq \max_{t \in T} D_{s[t,\,t(s)]}(\psi).$$

To prove this inequality, let t be any treatment and observe that a necessary condition for $\delta(t, \psi) > 0$ is that $m_t(\psi) \geq m_{t(s)}(\psi)$. For any t such that $\delta(t, \psi) > 0$,

$$\mu_{st(s)} - \mu_{st} \leq [m_t(\psi) - m_{t(s)}(\psi)] - [\mu_{st} - \mu_{st(s)}] = D_{s[t,\,t(s)]}(\psi) \leq \max_{t' \in T} D_{s[t',\,t(s)]}(\psi).$$

Given that $\delta(t, \psi) \geq 0$ for all t and that $\sum_{t \in T} \delta(t, \psi) = 1$, it follows that

$$U^*(P_s) - U(\delta, P_s, \psi) = \sum_{t:\,\delta(t,\,\psi) > 0} \delta(t, \psi)(\mu_{st(s)} - \mu_{st}) \leq \max_{t \in T} D_{s[t,\,t(s)]}(\psi).$$

II: Each variable $D_{s[t,\,t(s)]}(\psi)$ is a sum of independent mean zero variables

$$D_{s[t,\,t(s)]}(\psi) = [m_t(\psi) - \mu_{st}] - [m_{t(s)}(\psi) - \mu_{st(s)}] = \sum_{j \in N(t)} (u_j - \mu_{st})/n_t - \sum_{j \in N[t(s)]} (u_j - \mu_{st(s)})/n_{t(s)}.$$

Inequality (4.16) of Hoeffding (33) applies to each element of both sums on the right-hand side. This inequality shows that, for any $c > 0$,

$$\mathrm{E}_s\{\exp\{c[(u - \mu_{st})/n_t)]\} \leq \exp[c^2 n_t^{-2} M^2/8]$$

for each element of the first sum and

$$\mathrm{E}_s\{\exp\{c[(u - \mu_{st(s)})/n_{t(s)})]\} \leq \exp[c^2 n_{t(s)}^{-2} M^2/8]$$

for each element of the second sum. The statistical independence of these elements implies that

$$\mathrm{E}_s\{\exp[c \cdot D_{s[t,\,t(s)]}(\psi)]\} \leq \exp[c^2 (n_t^{-1} + n_{t(s)}^{-1}) M^2/8].$$

III: The conclusion to Part I implies that the regret of δ in state s is bounded above as follows:

$$U^*(P_s) - W(\delta, P_s, Q_s) \leq \mathrm{E}_s[\max_{t \in T} D_{s[t,\,t(s)]}(\psi)].$$

We use the conclusion to Part II and a proof similar to Lemma 1.3 of Lugosi (34) to obtain an upper bound on $\mathrm{E}_s[\max_{t \in T} D_{s[t,\,t(s)]}(\psi)]$. For any $c > 0$, by Jensen's inequality,

$$\exp\{c \cdot \mathrm{E}_s[\max_{t \in T} D_{s[t,\,t(s)]}(\psi)]\} \leq \mathrm{E}_s\{\exp[c \cdot \max_{t \in T} D_{s[t,\,t(s)]}(\psi)]\} =$$

$$= \mathrm{E}_s\{\max_{t \in T} \exp\{c \cdot D_{s[t,\, t(s)]}(\psi)\}\} \;\le\; \mathrm{E}_s\{\sum_{t \in T} \exp[c \cdot D_{s[t,\, t(s)]}(\psi)]\} \;=$$

$$= 1 + \sum_{t \neq t(s)} \mathrm{E}_s\{\exp[c \cdot D_{s[t,\, t(s)]}(\psi)]\} \;\le$$

$$\le 1 + \sum_{t \neq t(s)} \exp[c^2(n_t^{-1} + n_{t(s)}^{-1})M^2/8],$$

where the last inequality follows from the conclusion to Part II. Taking the logarithm of both sides and dividing by c yields

$$\mathrm{E}_s[\max_{t \in T} D_{s[t,\, t(s)]}(\psi)] \;\le\; \frac{\ln\{1 + \sum_{t \neq t(s)} \exp[c^2(n_t^{-1} + n_{t(s)}^{-1})\, M^2/8]\}}{c} =$$

$$= N^{-1/2} M \, \frac{\ln\{1 + \sum_{t \neq t(s)} \exp[d^2(p_t^{-1} + p_{t(s)}^{-1})/8]\}}{d},$$

here $d = N^{-1/2}Mc$.

IV. The conclusion to III holds in every state s. Hence, the maximum regret of δ is bounded above by

$$\max_{s \in S} [U^*(P_s) - W(\delta, P_s, Q_s)] \;\le\; N^{-1/2} M \max_{s \in S} \frac{\ln\{1 + \sum_{t \neq t(s)} \exp[d^2(p_t^{-1} + p_{t(s)}^{-1})/8]\}}{d}.$$

The summation $\sum_{t \neq t(s)} \exp[d^2(p_t^{-1} + p_{t(s)}^{-1})/8]$ is maximized in a state s such that $t(s) = t^*$, where t^* is a treatment with the smallest sample size. This holds because $p_{t^{**}} \geq p_{t^*}$ for any $t^{**} \neq t^*$. Hence,

$$\sum_{t \neq t^*} \exp[d^2(p_t^{-1} + p_{t^*}^{-1})/8] \;-\; \sum_{t \neq t^{**}} \exp[d^2(p_t^{-1} + p_{t^{**}}^{-1})/8] \;=$$
$$= \sum_{t \neq t^*,\, t^{**}} \{\exp[d^2(p_t^{-1} + p_{t^*)}^{-1})/8] - \exp[d^2(p_t^{-1} + p_{t^{**}}^{-1})/8]\} \;\ge\; 0.$$

The above shows that

$$\max_{s \in S} [U^*(P_s) - W(\delta, P_s, Q_s)] \;\le\; N^{-1/2} M \frac{\ln\{1 + \sum_{t \neq t^*} \exp[d^2(p_t^{-1} + p_{t^*}^{-1})/8]\}}{d}.$$

Finally, observe that the above inequality holds for all $d > 0$. This yields result (S7).

V. If $n_t = n$ for all t, then $p_t = n/N$ for all t. It follows that

$$1 + \sum_{t \neq t^*} \exp[d^2(p_t^{-1} + p_{t^*}^{-1})/8] \;\le\; K \cdot \exp[(n/N)^{-1} d^2/4].$$

Hence, (S7) implies that

$$\max_{s \in S} [U^*(P_s) - W(\delta, P_s, Q_s)] \le N^{-1/2} M \min_{d>0} \frac{\ln\{K \cdot \exp[(n/N)^{-1} d^2/4]\}}{d}$$

$$= n^{-1/2} M \min_{h>0} \frac{\ln[K \cdot \exp(h^2/4)]}{h},$$

where $h = (n/N)^{-1/2} d$. The minimum is obtained at $h = 2(\ln K)^{1/2}$. This implies result (S8).

Q. E. D.

Proposition 3: Consider any positive integer n. Among all designs with total sample size $K \cdot n$,
(a) bound (S6) in Proposition 1 is minimized by a balanced design with $n_t = n$ for all t.
(b) bound (S7) in Proposition 2 is minimized by a balanced design with $p_t = 1/K$ for all t. □

Proof:
a) Bound (S6) of Proposition 1 established that maximum regret is less than

$$\tfrac{1}{2} e^{-1/2} M \sum_{t \neq t^*} (n_t^{-1} + n_{t^*}^{-1})^{1/2}.$$

For a balanced design, the sum in the bound equals

$$\sum_{t \neq t^*} (n^{-1} + n^{-1})^{1/2} = (K-1) 2^{1/2} n^{-1/2}.$$

For any design with $\sum_{t \in T} n_t = K \cdot n$, the minimum sample size is $n_{t^*} \le n$. This and the fact that $K \ge 2$ imply that

(S9) $$\sum_{t \neq t^*} (n_t + n_{t^*}) = K \cdot n + (K-2) n_{t^*} \le 2(K-1) \cdot n.$$

Applying Jensen's inequality to $f(x) = x^{-1}$, which is convex for $x > 0$, yields $n_t^{-1} + n_{t^*}^{-1} \ge 4(n_t + n_{t^*})^{-1}$. In the derivation below, we apply this inequality to the sum in bound (S6), then apply Jensen's inequality to the function $f(x) = x^{-1/2}$, which is convex for $x > 0$, and then combine inequality (S9) with the fact that $f(x) = x^{-1/2}$ is a decreasing function:

$$\begin{aligned} \sum_{t \neq t^*} (n_t^{-1} + n_{t^*}^{-1})^{1/2} &\ge \sum_{t \neq t^*} [4(n_t + n_{t^*})^{-1}]^{1/2} = 2 \sum_{t \neq t^*} (n_t + n_{t^*})^{-1/2} \ge \\ &\ge 2(K-1) \cdot [(K-1)^{-1} \sum_{t \neq t^*} (n_t + n_{t^*})]^{-1/2} \ge \\ &\ge 2(K-1) \cdot [(K-1)^{-1} 2(K-1) \cdot n]^{-1/2} = (K-1) 2^{1/2} n^{-1/2}. \end{aligned}$$

This shows that the bound for any design with total sample size $K \cdot n$ is no smaller than the bound with a balanced design.

(b) Bound (S7) of Proposition 2 established that maximum regret is less than

$$N^{-1/2} M \min_{d>0} \frac{\ln\{1 + \sum_{t \neq t^*} \exp[d^2 (p_t^{-1} + p_{t^*}^{-1})/8]\}}{d}.$$

For a balanced design and any $d > 0$, the sum in the bound equals

$$\sum_{t \neq t^*} \exp[d^2(p_t^{-1} + p_{t^*}^{-1})/8] = \sum_{t \neq t^*} \exp\{d^2[(1/K)^{-1} + (1/K)^{-1}]/8\} = (K-1)\exp(d^2K/4).$$

We will show that for any $(p_t, t \in T)$ such that $\sum_{t \in T} p_t = 1$,

$$\sum_{t \neq t^*} \exp[d^2(p_t^{-1} + p_{t^*}^{-1})/8] \geq (K-1)\exp(d^2K/4).$$

This result, which holds for all $d > 0$, and the fact that $\ln(\cdot)$ is an increasing function show that the bound for any design with total sample size $K{\cdot}n$ is no smaller than the bound with a balanced design.

Applying Jensen's inequality to $f(x) = x^{-1}$, which is convex for $x > 0$, yields $p_t^{-1} + p_{t^*}^{-1} \geq 4(p_t + p_{t^*})^{-1}$. Given that $\exp(\cdot)$ is increasing, it follows that

$$\text{(S10)}\ \sum_{t \neq t^*} \exp[d^2(p_t^{-1} + p_{t^*}^{-1})/8] \geq \sum_{t \neq t^*} \exp[(d^2/2)\cdot(p_t + p_{t^*})^{-1}].$$

Applying Jensen's inequality to the convex function $f(x) = \exp(x)$ yields

$$\text{(S11)}\ \sum_{t \neq t^*} \exp[(d^2/2)\cdot(p_t + p_{t^*})^{-1}] \geq (K-1)\exp\{(K-1)^{-1}\sum_{t \neq t^*}[(d^2/2)\cdot(p_t + p_{t^*})^{-1}]\} =$$

$$= (K-1)\exp[(d^2/2)\cdot(K-1)^{-1}\sum_{t \neq t^*}(p_t + p_{t^*})^{-1}].$$

Applying Jensen's inequality to $f(x) = x^{-1}$, which is convex for $x > 0$, yields

$$(K-1)^{-1}\sum_{t \neq t^*}(p_t + p_{t^*})^{-1} \geq [(K-1)^{-1}\sum_{t \neq t^*}(p_t + p_{t^*})]^{-1} = \{(K-1)^{-1}[1 + (K-2)p_{t^*}]\}^{-1}.$$

Given that $K - 2 \geq 0$ and $p_{t^*} \leq 1/K$, it follows that $1 + (K-2)p_{t^*} \leq 2(K-1)/K$. Given that $f(x) = x^{-1}$ is a decreasing function, it follows that

$$\text{(S12)}\ (K-1)^{-1}\sum_{t \neq t^*}(p_t + p_{t^*})^{-1} \geq \{(K-1)^{-1}[2(K-1)/K]\}^{-1} = K/2.$$

Combining (S10), (S11), and (S12) with the monotonicity of $\exp(\cdot)$ yields

$$\sum_{t \neq t^*} \exp[d^2(p_t^{-1} + p_{t^*}^{-1})/8] \geq (K-1)\exp[d^2K/4].$$

Q. E. D.

Table S1 presents the values of bounds (S6), (S7), and (S8) for balanced designs. The three bounds vary identically with $Mn^{-½}$ but differently with the number of treatments. Proposition 1 provides a better bound for $K \leq 3$, while Proposition 2 provides a better bound for $K \geq 4$. Bound (S8) of Proposition 2 is simpler to compute than bound (S7) and is only marginally larger.

We have also computed bounds (S6) and (S7) for various unbalanced designs. We again find that bound (S6) is better for $K \leq 3$ and bound (S7) is better for $K \geq 4$. These results are not shown in the table.

Propositions 1 and 2 imply sufficient conditions on sample sizes for ε-optimality of ES rules. If the upper bound on maximum regret with a specified trial design is less than or equal to ε, then ES rules are ε-optimal with this design.

The findings are particularly simple with balanced designs. Then bound (S6) of Proposition 1 implies that an ES rule is ε-optimal if $n \geq (2e)^{-1}(K-1)^2(M/\varepsilon)^2$. Bound (S8) of Proposition 2 implies that an ES rule is ε-

optimal if $n \geq \ln K \cdot (M/\varepsilon)^2$. Table S1 gives the threshold sample size for bound (S7) of Proposition 2 for $K \leq 7$, which is $(M/\varepsilon)^2$ times the square of the relevant constant shown in the table.

It is important to bear in mind that Propositions 1 and 2 only imply simple sufficient conditions on sample sizes for ε-optimality of ES rules, not necessary ones. Proposition 1, for example, could be sharpened for balanced designs by replacing Hoeffding's inequality by Theorem 1.2 of Bentkus (35) and further improvements should be possible. The Bentkus inequality is expressed in terms of a tail probability of a binomial distribution and the resulting regret bound has to be evaluated numerically for each n. For large values of n, the regret bound could be up to 23.5% smaller than (S6).

In general it is difficult to compute the exact maximum regret of ES rules, hence difficult to determine how conservative the propositions are. An exception occurs when there are two treatments and outcomes are binary. Then the maximum regret of various decision rules can be computed numerically without large deviations bounds.

S2.3. ε-Optimality of Empirical Success Rules with Observable Covariates

The above analysis has assumed that members of the population have no observable covariates that may be used to condition treatment choice. Suppose now that persons have observable covariates taking values in a finite set X and that the planner can execute a trial with (treatment, covariate)-specific sample sizes $[n_{t\xi}, (t, \xi) \in T \times X]$. We consider the ES rule defined in Section S1.3, which assigns all persons with covariates ξ to treatments that maximize $m_{t\xi}(\psi)$ over T, $m_{t\xi}(\psi)$ being the average outcome in sub-sample $N(t, \xi)$.

There are at least two reasonable ways that a planner may wish to evaluate ε-optimality in this setting. First, he may want to achieve ε-optimality within each covariate group. This interpretation requires no new analysis. The planner should simply define each covariate group to be a separate population of interest and then apply the analysis of Section S2.2 to each group. The design that achieves group-specific ε-optimality with minimum total sample size equalizes sample sizes across groups.

Alternatively, the planner may want to achieve ε-optimality within the overall population, without requiring that it be achieved within each covariate group. This is the interpretation given in Section S1.3, when we defined rule δ to be ε-optimal if $W(\delta, P_s, Q_s) \geq U^*(P_s) - \varepsilon$ for all $s \in S$. In this case, the design that achieves ε-optimality with minimum total sample size does not equalize sample sizes across groups.

Propositions 1 and 2 easily extend to provide sample sizes sufficiently large to yield the latter interpretation of ε-optimality. Applying Proposition 1 to each group and aggregating the bounds across groups implies that the maximum regret of ES rules is bounded above by

$$\text{(S6')} \quad \max_{s \in S} [U^*(P_s) - W(\delta, P_s, Q_s)] \leq \tfrac{1}{2} e^{-\frac{1}{2}} M \sum_{\xi \in X} P(x = \xi) \sum_{t \neq t^*(\xi)} (n_{t\xi}^{-1} + n_{t^*(\xi)\xi}^{-1})^{\frac{1}{2}},$$

where $t^*(\xi) \in \operatorname{argmin}_{t \in T} n_{t\xi}$ denotes a treatment with the smallest sample size among individuals with covariate value ξ. When the design is balanced across treatments for each covariate, with $n_{t\xi} = n_\xi$ for all t, the bound is

$$\text{(S13)} \quad (2e)^{-\frac{1}{2}} (K - 1) M \sum_{\xi \in X} P(x = \xi) n_\xi^{-\frac{1}{2}}.$$

The analogous extension of Proposition 2 yields

$$\text{(S7')} \quad \max_{s \in S} [U^*(P_s) - W(\delta, P_s, Q_s)] \leq$$

$$\leq M \sum_{\xi \in X} P(x = \xi)\, N_\xi^{-½} \min_{d>0} \frac{\ln\{1 + \sum_{t \neq t^*(\xi)} \exp[d^2(p_{t\xi}^{-1} + p_{t^*(\xi)\xi}^{-1})/8]\}}{d},$$

where $N_\xi \equiv \sum_{t \in T} n_{t\xi}$ is total sample size for individuals with covariate value ξ and $p_{t\xi} \equiv n_{t\xi}/N_\xi$. When the design is balanced across treatments for each covariate, $N_\xi^{-½} = K^{-½} n_\xi^{-½}$, $p_{t\xi} = 1/K$ for all (t, ξ), and the bound in (S7') simplifies to

$$\text{(S14)}\; K^{-½} \min_{d>0} \frac{\ln\{1 + (K-1)\exp(d^2K/4)\}}{d} \; M \sum_{\xi \in X} P(x = \xi)\cdot n_\xi^{-½}.$$

Bounds (S13) and (S14) can easily be evaluated for any candidate treatment-balanced design to verify whether it suffices to enable ε-optimal treatment rules. The constants preceding $\sum_{\xi \in X} P(x = \xi)\cdot n_\xi^{-½}$ in these bounds are given in Table S1 for $K \leq 7$.

Given a predetermined maximum total sample size N, minimizing bounds (S13) and (S14) is achieved by choosing $(n_\xi, \xi \in X)$ to minimize $\sum_{\xi \in X} P(x = \xi)\cdot n_\xi^{-½}$ subject to the constraint $\sum_{\xi \in X} n_\xi \leq N/K$. Given that the objective function is decreasing in each n_ξ, the constraint binds. The Lagrangian expression of the constrained minimization problem is

$$\text{(S15)}\; \mathrm{L}[(n_\xi, \xi \in X), \lambda] \equiv \sum_{\xi \in X} P(x = \xi)\cdot n_\xi^{-½} + \lambda\left(\sum_{\xi \in X} n_\xi - N/K\right).$$

A simple approximation to the minimization problem results if one treats $(n_\xi, \xi \in X)$ as continuous variables rather than as integer sample sizes. Then the first order conditions for minimization of $\mathrm{L}(\cdot, \cdot)$ yield

$$\text{(S16)}\; -½P(x = \xi)\cdot n_\xi^{-3/2} + \lambda = 0, \text{ all } \xi \in X.$$

This implies that $n_\xi = (2\lambda)^{-⅔}P(x = \xi)^{⅔}$. It follows that, to solve problem (S16), the relative sample sizes for any pair (ξ, ξ') of covariate values have the approximate ratio

$$\text{(S17)}\; n_\xi/n_{\xi'} = [P(x = \xi)/P(x = \xi')]^{⅔}.$$

For the case when the covariate takes two values, a similar result is obtained by Schlag (24).

A planner who uses (S17) to choose the trial design makes the covariate-specific sample size increase with the prevalence of the covariate group in the population, albeit less than proportionately. Covariate-specific maximum regret commensurately decreases with the prevalence of the covariate group.

Table S1: Bounds in Propositions 1 and 2 for Balanced Designs with n Subjects per Treatment

K =	2	3	4	5	6	7	
Bound (S6)	0.4289	0.8578	1.2866	1.7155	2.1444	2.5733	$\cdot Mn^{-½}$
Bound (S7)	0.6539	0.9279	1.0892	1.1999	1.2827	1.3481	$\cdot Mn^{-½}$
Bound (S8)	0.8326	1.0481	1.1774	1.2686	1.3386	1.3950	$\cdot Mn^{-½}$